DETERMINING THE ACCURACY IN SUPERVISED FUZZY CLASSIFICATION PROBLEMS

DANIEL GOMEZ

Faculty of Statistics, Complutense University
Madrid, Spain
dagomez@estad.ucm.es

JAVIER MONTERO

Faculty of Mathematics, Complutense University
Madrid, Spain
monty@mat.ucm.es

A large number of accuracy measures for image classification are actually available in the literature for cris classification. Overall accuracy, producer accuracy, user accuracy, kappa index and tau value are some examples. But in contrast to this effort in measuring the accuracy in a crisp framework, few proposals can be found in order to determine accuracy for soft classifiers. In this paper we define some accuracy measures for soft classification that extend some classical accuracy measures for crisp classifiers. This class of measures takes into account the preferences of the decision maker in order to differentiate some errors that in practice may not be have same relevance.

1. Introduction

No supervised classification is complete until an assessment of its accuracy has been performed. Although this problem has been addressed by many researches, most of them assume that both classifier and expert are crisp. And most fuzzy approaches recently published are still based on the confusion matrix. The pioneer work is due to Binaghi [1] et al., who built a *fuzzy matrix error* which generalizes the error matrix within a remote seing classification problem: for each object p and each cell (i, j), they determine (based on the *min* operator) the degree to which p has been classified in class j by the expert and in class i by the classifier. After that, this information is aggregated for each object p to obtain the *fuzzy error matrix* (see [1] for more details). Although this fuzzy error matrix presents some advantages compared with a standard classical approach, some misbehavior

appears when the fuzzy classification that is evaluated is not a Ruspini partition, something that will happen too often (see [5]). In order to solve this problem, Gómez *et al.* [8] defined a new family of disagreement weighted measures that extend the most popular accuracy measures in classification: the overall and the Kappa statistic for classical *hard* (crisp) classifications. Moreover, such an alternative weighted accuracy measure we can avoid the assumption of equally important errors (a standard assumption within fuzzy accuracy assessments). A key problem then is the determination of the importance of each error. In this work, we present different alternatives that take into account decision maker preferences.

2. Measuring the errors

From a mathematical point of view, an object p that has been classified by the expert (E) or by a classifier (C) into class i, can be modelled as a k dimensional vector $(0, \ldots, 1, \ldots, 0)$, k being the number of different classes under consideration. We will denote by P the set of objects that has to be classified and by $T \subset P$ the training set, necessary in any supervised classification algorithm. In order to extend the concept of error between an expert (E) (reference data) and the classifier (C) let us introduce the following definition.

Definition 2.1. Given a set of objects P and a family of classes $A_1 \ldots A_k$ under consideration, E the expert function and C the classifier function, then the error D of the object p given by the classifier C is defined as:

$$D(E(p), C(p), p) = Min \left\{ 1, \sum_{j=1}^{k} w_{ij} |E(p)_j - C(p)_j| \right\}$$

where $E(p)_j$ is the j-th coordinate of $E(p)$, $C(p)_j$ is the j-th coordinate of the classifier function $C(p)$, i represents the class to which p is assigned the largest degree of membership $Max\left\{(E(p))_{1 \leq r \leq k}\right\} = (E(p))_i$ and each $w_{ij} \in \Re$ represents the importance of the error when an object that belongs to class i is classified into class j.

Notice that the above definition requires that the maximum in the $E(p)$ vector is unique. In the case in which the maximum is reached in more than one component we will take the average of the different errors between these classes. So, if we have, for example, $E(p) = (0.4, 0.4, 0.2)$ and $C(p) = (0.4, 0.4, 0.3)$ two different disagreement measures (depending on where the

maximum is reached) are defined in this example as: $0.1w_{13}$ if we take A_1 as the maximum and $0.1w_{23}$ if we take A_2 as the maximum. For this example, the definition for disagreement D that we propose is the average, that is $0.05w_{13} + 0.05w_{23}$. Taking into account this, the final disagreement will be the $Min\{1, 0.05w_{13} + 0.05w_{23}\}$. More generally, importance errors w_{ij} may depend on the whole vector $E(p)$, so its dispersion can be taken into account.

Let us note that if all errors are considered equally important ($w_{ij} = 1$ for all $i \neq j$ and $w_{ii} = 0$ for all i), and both classifier and expert are crisp, then the error function above defined coincides with the classical approach, i.e.

$$D(E, C, p) = \begin{cases} 0 \; if & E(p) = C(p) \\ 1 \; if & E(p) \neq C(p) \end{cases}$$

3. Accuracy measures

From now on we will define the agreement measure between expert and classifier as $A(E, C, p) = 1 - D(E, C, p)$. Once the error (agreement) function is obtained and the weights are determined, the overall accuracy and the kappa index can be obtained by means of an adequate aggregation of errors for each object.

Definition 3.1. Given P the object set, $A_1 \dots A_k$ the family of classes under consideration, E the expert function and C the classifier function, we define the overall accuracy (O^C) as:

$$O^C = \sum_{p \in T} \frac{1 - D(E, C, p)}{|T|} = \sum_{p \in T} \frac{A(E, C, p)}{|T|}$$

Let us note that if the classifier produces a Ruspini's partition (i.e., $\sum_{i=1}^{k} C_i(p) = 1, \; \forall p \in T$), and the expert is crisp, then the overall accuracy measure above defined coincides with the overall accuracy defined by Binaghi [1] *et al.* In a more general case, Ruspini's assumption is not fulfilled and then Binaghi's approach may produce strange results, as shown below.

Example 3.1. Let us suppose that all errors are considered equally important, $E(p) = (0.2, 0.1, 0.2)$ and $C(p) = (0.4, 0.3, 0.2)$ for an object $p \in T$. If we build the fuzzy error matrix defined in [1], we obtain:

$$X = \begin{pmatrix} 0.2\ 0.2\ 0.2 \\ 0.1\ 0.1\ 0.1 \\ 0.2\ 0.2\ 0.2 \end{pmatrix}$$

In this matrix, $x_{ij} = min\{E(p)_i, C(p)_j\}$. Following this fuzzy error matrix, the overall accuracy defined in [1] is $\frac{\sum_i x_{ii}}{\sum_i E(p)_i} = \frac{0.2+0.1+0.2}{0.2+0.1+0.2} = 1$. So, Binaghi's approach suggests a perfect agreement between the expert and the classifier. But in our opinion this is inappropriate. On the contrary, our agreement measure will suggest a more accurate difference between expert and classifier: $A(E, C, p) = 0.8$.

Example 3.2. Let us suppose that all errors are considered equally important, $E_1(p) = (0.4, 0.1, 0.2, 0.3)$, $E_2(p) = (0.4, 0.3, 0, 0)$, $E_3(p) = (0.4, 0, 0.1, 0.3)$, $E_4(p) = (0.4, 0, 0.2, 0)$ and $C(p) = (1, 0, 0, 0)$ for a given object $p \in T$. In the Binaghi case, the overall accuracy can be assigned an overall accuracy of 1 in all four cases. But again this result is not appropriate, and in fact our agreement measure establishes differences between expert and classifier: $A(E_1, C, p) = 0.4$, $A(E_2, C, p) = 0.7$, $A(E_3, C, p) = 0.6$ and $A(E_4, C, p) = 0.8$.

An *Extended Kappa statistic* is next proposed, based on the previous Kappa statistic but allowing comparisons between arbitrary classifiers (a crisp classifier with a crisp data reference set and equal weights, a crisp classifier with a crisp data reference set and non-equal weights, a fuzzy classifier with a crisp data reference set and equal weights, and a fuzzy classifier with a crisp data reference set and non-equal weights). It is important to note that this new definition is an extension of the standard Kappa measure for two raters.

Definition 3.2. Given P the set of objects, $T \subset P$ the accuracy data set with cardinality t, $A_1 \dots A_k$ the family of classes under consideration, E the expert function and C the classifier function, we define the Extended Kappa statistic K_E as:

$$K_E = \frac{\hat{p_o} - \hat{p_c}}{1 - \hat{p_c}}$$

where $\hat{p_o} = O^C$ is the overall accuracy

$$\hat{p_c} = \sum_{p \in T} \sum_{q \in T} \frac{1 - D\left(C(p), E(q)\right)}{t^2}$$

where

$$D\left(C(p),E(q)\right)=Min\left\{1,\sum_{j=1}^{k}w_{ij}|\left(C(p)\right)_j-\left(E(q)\right)_j|\right\}$$

with $Max\left\{\left(E(q)\right)_{1\leq r\leq k}\right\}=\left(E(q)\right)_i$.

4. Obtaining weights.

As it can be perceived from the disagreement measure given in definition 2.1, the weights that represent the importance of the different errors play an extremely important role. In the following two subsections we propose two alternative techniques in order to determine the importance of errors. The first one is based on a multi-criteria decision making approach, and the second one is based on the distance between fuzzy sets.

4.1. *A multi-criteria approach*

It is a standard assumption in accuracy assessment that all errors are equally important. Introducing weights to account the relative importance of errors will introduce in the system the opinion of the expert. As a consequence, a different weight matrix for each measure will be required. From a multi-criteria point of view there are several available approaches in order to determine weights (see, e.g., [7,10]). For example, if we want to determine the importance of each error based on Saaty methodology, we have to obtain first the Saaty matrix. Once the Saaty matrix has been defined, the weights are computed, for example, as the eigenvector associated to the maximum eigenvalue (see [10]). Of course, other alternatives can be proposed.

4.2. *Fuzzy distances*

In the framework of remote sensing classification problems, images can be described in fuzzy terms by means of the spectral features of each class. Consequently, for each class A_i, and for each band B_r, we have the functions $\mu_{A_k}^{B_r}$. Taking into account that for each A_j we have $\left(\mu_{A_j}^{B1},\ldots,\mu_{A_j}^{Bm}\right)$, a distance function between fuzzy sets could be applied for each pair of classes, $D(A_i,A_j)=d_{ij}$. On one hand, small distances d_{ij} represent a hight similitude between classes, so error will not be relevant. On the other hand, high values of d_{ij} represent different classes or big errors. Taking into account this information, the weights matrix could be calculated proportionally to distance values.

5. Final comments

As a final comment we want to stress the relevance of our proposal, since very few accuracy measures are available for fuzzy classification. If offering a quantitative measure of the quality of every classification is an essential objective within a crisp framework, such measures will play a more relevant role under fuzziness, where certain visual arguments are much more difficult to argue.

Acknowledgements

This work has been partially supported by the National Science Foundation of Spain, grant TIN2006-06290.

References

1. Binaghi, E., Brivio, P.A., Ghezzi, P. & Rampini, A. (1999). A fuzzy set based accuracy assessment of soft classification. *Pattern Recognition Letters*, 20, 935948
2. Cohen, J. (1960). A coeficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46
3. Cohen, J. (1968). Weighted Kappa: Nominal Scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220
4. Congalton, R.G. & Green K. (1999). *Assessing the accuracy of remote sensed data: Principles and Practices*. London New York and Washinton D.C: Lewis publishers
5. Del Amo, A., Montero, J., Biging, G. & Cutello, V. (2004): Fuzzy classification systems. *European Journal of Operational Research* 156:459–507.
6. Del Amo, A., Gómez, D., Montero, J. & Biging, G. (2001): Relevance and redundancy in fuzzy classification systems. *Mathware and Soft Computing* 8:203–216.
7. Gonzalez-Pachón, J., Gómez, D., Montero, J., & Yánez, J. (2003). Soft dimension theory. *Fuzzy set and Systems*, 137, 137-149
8. Gómez D., Biging G., and J. Montero. Accuracy statistics for judging soft classification. *International Journal of Remote Sensing* DOI.10.1080/01431160701311325
9. Ruspini, E.H. (1969). A new approach to clustering. *Information and Control*, 15, 22-32
10. Saaty, T.L. (1994): *Fundamentals of Decision Making with the Analytic Hierarchy Process*. RWS Publications, Pittsburgh (Revised in 2000).
11. Uebersax, J.S. (1982). A generalized Kappa coefficient. *Educational and Psychological Measurement*, 42, 181-183